Prepared for submission to JHEP CP3-15-01

Playing Tag with ANN: Boosted Top Identification with Pattern Recognition

Leandro G. Almeida,[a] **Mihailo Backović,**[b] **Mathieu Cliche,**[c] **Seung J. Lee,**[d,e] **Maxim Perelstein**[c]

[a]*Institut de Biologie de l'École Normale Supérieure (IBENS), Inserm 1024- CNRS 8197, 46 rue d'Ulm, 75005 Paris, France*

[b]*Center for Cosmology, Particle Physics and Phenomenology - CP3, Universite Catholique de Louvain, Louvain-la-neuve, Belgium*

[c]*Laboratory for Elementary Particle Physics, Cornell University, Ithaca, NY 14853, USA*

[d]*Department of Physics, Korea Advanced Institute of Science and Technology, 335 Gwahak-ro, Yuseong-gu, Daejeon 305-701, Korea*

[e]*School of Physics, Korea Institute for Advanced Study, Seoul 130-722, Korea*

E-mail: almeida@biologie.ens.fr, mihailo.backovic@uclouvain.be, mc863@cornell.edu, sjjlee@kaist.ac.kr, mp325@cornell.edu

Abstract: Many searches for physics beyond the Standard Model at the Large Hadron Collider (LHC) rely on top tagging algorithms, which discriminate between boosted hadronic top quarks and the much more common jets initiated by light quarks and gluons. We note that the hadronic calorimeter (HCAL) effectively takes a "digital image" of each jet, with pixel intensities given by energy deposits in individual HCAL cells. Viewed in this way, top tagging becomes a canonical pattern recognition problem. With this motivation, we present a novel top tagging algorithm based on an Artificial Neural Network (ANN), one of the most popular approaches to pattern recognition. The ANN is trained on a large sample of boosted tops and light quark/gluon jets, and is then applied to independent test samples. The ANN tagger demonstrated excellent performance in a Monte Carlo study: for example, for jets with p_T in the $1100 - 1200$ GeV range, 60% top-tag efficiency can be achieved with a 4% mis-tag rate. We discuss the physical features of the jets identified by the ANN tagger as the most important for classification, as well as correlations between the ANN tagger and some of the familiar top-tagging observables and algorithms.

Contents

1 Introduction

Many extensions of the Standard Model (SM) predict new particles with masses around the TeV scale. Searches for such new particles form a major component of the experimental program at the Large Hadron Collider (LHC). In most models, the new particles are unstable, and their decays often contain weak-scale SM states, namely the W and Z bosons, the Higgs boson, and the top quark. Searches for final states containing top quarks are particularly important, due to the special role played by the top sector in many models of electroweak symmetry breaking. Decays of heavy new particles with mass above the electroweak scale typically result in highly energetic, relativistic top quarks in the lab frame. Identifying and characterizing such "boosted" top quarks in the data is crucial for new physics searches and tests of naturalness [1] at the LHC, especially as the bounds on the new physics mass scales in many candidate models are pushed higher. Examples of new physics leading to boosted top signatures include Kaluza-Klein gluons [2, 3] and string Regge states [4] of the Randall-Sundrum model, stops [5] and gluinos [6] of supersymmetry, top and light quark partner decays in Composite Higgs models [7–12], and many others.

Due to relativistic kinematics, the decay products of a boosted top quark are highly collimated. For instance, hadronic decay of a top quark of $p_T \sim 1$ TeV would produce three quarks collimated into a cone of rough size $R \sim 0.4$ and result in a specific pattern of hadronic activity in the detector. Classical event reconstruction techniques are inadequate to tag and measure such topologies, as most of the showered radiation falls into a small angular region. One solution is to cluster the event with a large jet cone ($R \sim 1$), and consider the features of energy distribution inside such "fat" jets (so-called jet substructure), instead of correlations between individual small radius jets. Over the past decade, a variety of methods for boosted top tagging via jet substructure have been developed (see Ref. [13]

for a review), most of which can be cast into several (non exclusive) groups. Jet shapes are observables based on various moments of the jet energy distribution. Notable examples are angular correlations studied extensively in Ref. [14], sphericity tensors [15, 16] and other perturbatively calculable jet shapes [17]. Considerations of jet clustering history led to development of numerous Filtering jet substructure methods [18–20], where the differences in the late steps of jet clustering between heavy SM states and QCD jets from light partons have been successfully applied in tagging of heavy SM states. Furthermore, Prong Taggers such as N-subjettiness [21, 22] exploit the differences in the number of hard energy depositions within the boosted jet (e.g. three-body top decays compared to the typical two-body splitting of a light jet). Parton level models of boosted decays and kinematic constraints built into them can also be used to study jet substructure, with the Template Overlap Method (TOM) [23–26] being the most notable example. More recently, Matrix Element Method [27, 28] inspired techniques such as Shower Deconstruction have emerged [29, 30], where a boosted jet is tagged using approximations to hard matrix elements and the parton shower. Soft drop declustering (a generalization of modified mass drop tagging) is another method which has been recently developed for removing non-global contributions (soft radiation) to the jet [31]. Several of these methods have been implemented in the analyses of the LHC data by the CMS and ATLAS collaborations; see, for example Ref. [32–35].

In this paper, we pursue an alternative approach to jet substructure. Experimentally, information about hadronic activity in an event comes mainly from the hadronic calorimeter (HCAL), with the basic observable being the energy deposited in each of the HCAL cells. One can think of the information provided by the HCAL as a *digital image*, with each cell (or topo-cluster) being identified as a pixel, and with energy deposit in the cell corresponding to the intensity (or grayscale color) of that pixel. From this point of view, boosted top identification is simply a classic image-recognition problem: distinguishing the energy-deposit patterns characteristic of boosted tops from patterns due to other sources, such as the usual QCD jets. This suggests that computational algorithms developed in the field of image recognition could be of use in boosted top tagging.[1]

With this motivation, we constructed a new top tagger algorithm based on one of the most popular approaches to image recognition, Artificial Neural Networks (ANNs). In this approach, each jet is classified as top or non-top according to a highly non-linear scoring function. The function contains multiple adjustable parameters, called weights. These are chosen using a training procedure, in which the ANN is presented with a large sample of jets that are known to be top or non-top, and the weights are chosen to maximize the number of correctly identified jets in this sample. (In our study, all samples are generated by Monte Carlo simulations. In experimental applications, ANN may be trained on either MC samples or carefully selected "calibration" data sets.) Having fixed the weights, the

[1]Recently, Ref. [36] studied jet substructure as an image recognition problem in the context of boosted W tagging as well gluon/quark discrimination. The authors utilised a linear Fisher discriminant trained on a sample of signal and background events, in order to distinguish the desired events from the backgrounds. The method out-performs the existing methods of W tagging, illustrating the benefits of the image recognition approach to jet substructure.

ANN is then applied to independent samples containing both top and non-top jets, and asked to discriminate between them. We find that the performance of the ANN tagger significantly exceeds that of several popular tagging algorithms currently in use over a wide range of p_T, demonstrating the practical utility of this approach.

The paper is organized as follows. Section 2 describes the MC event samples used for training and testing the ANN tagger, as well as the pre-processing steps applied to these samples before the ANN is applied. Section 3 contains a detailed description of the ANN tagger, including the network architecture and the training algorithms we employed. In Section 4, we present the results of our study of ANN tagger performance and comparisons with other popular taggers. We also discuss the physical features of jets that are dominant in the ANN classification, and the extent to which ANN output is correlated with that of other taggers. We conclude with a recap and a brief discussion of directions for future research in Section 5. An Appendix contains a brief description of the top taggers we use for the purpose of comparison with the ANN tagger.

2 Event Generation and Pre-Processing

We generate benchmark event samples with `MadGraph 5` [37] at leading order, and shower them with `Pythia 6` [38]. In order to study the effects of different showering algorithms on the results, we also generate separate data samples showered with `Pythia 8` [39]. For simplicity, we extract a pure sample of top jets from a Standard Model top pair-production simulation, at leading order with no matching. The tops are decayed in `MadGraph 5`, so that the angular distribution of the decay products is modeled correctly. Similarly, we generate the light jet sample from a simulation of the QCD di-jet process, including both quarks and gluons in the final state, but no matching to extra jets. Fiducial cut $|\eta| \leq 5.0$ is imposed at the hadron level. We cluster the events using the `fastjet` [40] implementation of the anti-k_T algorithm [41] with a large jet cone of $R = 1.0$. For our analysis, we only use the highest p_T jet in each event, and impose the cut $|\eta_{\text{jet}}| \leq 2.5$. We consider samples of jets within three jet p_T ranges: 500 – 600 GeV, 800 – 900 GeV and 1100 – 1200 GeV. These three bins span a range of jet p_T values relevant for top tagging at the LHC, while analyzing them separately provides information about p_T sensitivity of the tagging efficiency and other parameters. Unless otherwise noted, we impose a cut on the jet mass (*i.e.* the invariant mass of all particles assigned to the jet), selecting jets within a window

$$130 \text{ GeV} < m_J^{R=1.0} < 210 \text{ GeV}. \tag{2.1}$$

A vast majority of top jets fall within this mass range, while most QCD jets are rejected by this cut. Discriminating the remaining QCD jets from top jets is the task for the top tagger.

In order to form an input to the ANN tagger, we preprocess each jet as follows. First, we find the center of the jet, defined by the sum of the coordinates of all particles weighted by their energies,

$$\eta_C = \frac{1}{E}\sum_j \eta_j E_j, \quad \phi_C = \frac{1}{E}\sum_j \phi_j E_j, \tag{2.2}$$

where $E = \sum_j E_j$ is the total energy of the jet. We then shift the coordinates of each particle so that the jet is centered at the origin in the new coordinates:

$$\eta'_j = \eta_j - \eta_C, \quad \phi'_j = \phi^j - \phi_C. \tag{2.3}$$

Further, we find the jet "principal axis" in the (η, ϕ) plane, defined by

$$\tan(\theta) = \frac{\sum_j \frac{\phi'_j \cdot E_j}{\Delta R'}}{\sum_j \frac{\eta'_j \cdot E_j}{\Delta R'}}, \quad \Delta R' = \sqrt{\eta'^2_j + \phi'^2_j}, \tag{2.4}$$

and rotate the coordinate system so that this principal axis is the same direction $(+\eta)$ for all jets:

$$\eta''_j = \eta'_j \cdot \cos(\theta) + \phi'_j \cdot \sin(\theta), \tag{2.5}$$

$$\phi''_j = -\eta'_j \cdot \sin(\theta) + \phi'_j \cdot \cos(\theta). \tag{2.6}$$

These coordinate transformations remove information about the jet position in the calorimeter and its orientation in the (η, ϕ) plane. Both pieces of information are irrelevant for top tagging, and removing them from consideration allows the ANN tagger to focus on the irreducible physical differences between top and QCD jets.[2]

In the new coordinates, nearly all (98%) of the particles assigned to a given jet fall within a window of $\eta'' \in [-\pi/2, \pi/2]$ and $\phi'' \in [\pi/2, \pi/2]$. We model the HCAL response to the jet by dividing this window into 30×30 square cells. (The cell size is approximately 0.1×0.1, close to the realistic values in ATLAS and CMS.) The normalized energy deposited in each cell, ε_{ab} $(a, b = 1 \ldots 30)$, is computed by adding up the energies of all particles falling within that cell, and dividing by the total energy of the jet. (The last step is once again necessary to remove information irrelevant for top tagging, in this case the total jet energy.) By construction, ε_{ab} is dimensionless and lies between 0 and 1. In the language of image processing, each jet has been converted into an image with 30×30 pixels, with a grayscale color of each pixel given by the corresponding ε_{ab}. These images can now be classified by an Artificial Neural Network (ANN), described in the following section.

3 ANN Tagger

ANN tagger is based on a feed-forward neural network with an input layer consisting of $30 \times 30 = 900$ nodes, one for each calorimeter cell; two hidden layers, of 100 nodes each, to process the signal; and an output layer consisting of a single node, whose value Y is interpreted as the probability that a given jet comes from a boosted top decay. The architecture of the network is shown in Fig. 1. (For pedagogical introduction to Artificial Neural

[2]As an exercise, we also attempted to train the neural network on a set of jets with randomly oriented principal axes, *i.e.* without the rotation (2.6). We found that this procedure still yields an effective tagger; presumably, the neural net learns to ignore the axis orientation information during the training process. However, to achieve the same tagging performance, the randomly-oriented training set needs to be significantly larger.

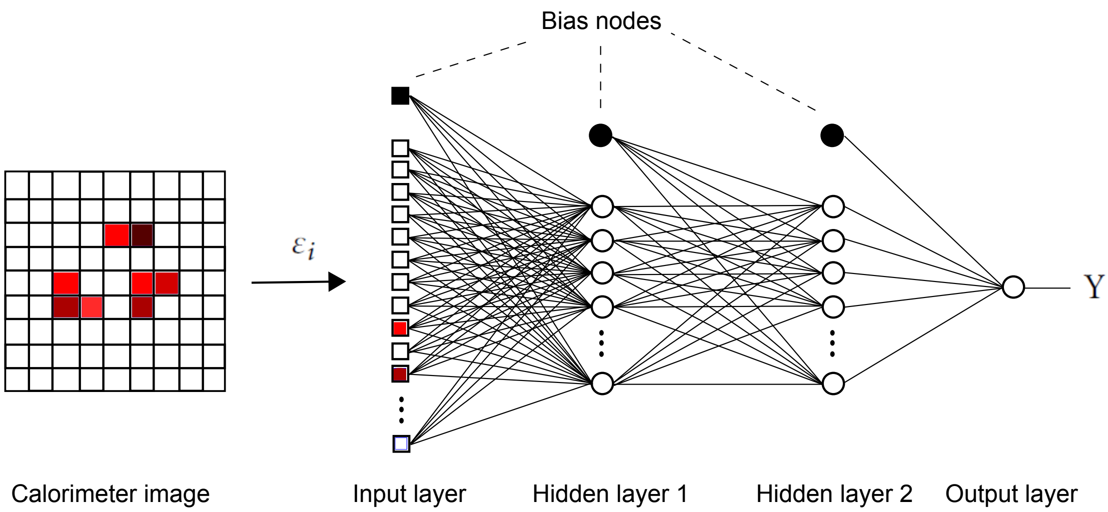

Figure 1. Graphical representation of the Artificial Neural Network (ANN).

Networks in the context of image recognition, see for example [42].) Mathematically, the ANN can be thought of as a succession of non-linear transformations:[3]

$$\epsilon_i \to h_i^{(1)} = f(W_{ij}^{(1)}\epsilon_j + b_i^{(1)}) \to \cdots \to h_i^{(l)} = f(W_{ij}^{(l)}h_j^{(l-1)} + b_i^{(l)}) \to Y = f(W_j^{(O)}h_j^{(l)} + b^{(O)}), \tag{3.1}$$

where f is the so-called activation function, chosen to be

$$f(z) = \frac{1}{1+e^{-z}}. \tag{3.2}$$

The inputs ϵ_i are simply the normalized energy deposits ε_{ab} defined above, rearranged in a single 900-dimensional vector: $\varepsilon_{ab} \equiv \epsilon_{30a+b}$. The weights $W_{ij}^{(L)}$ and the biases $b_i^{(L)}$ are numbers determined by the training procedure, which we will now describe.

To train the network, we use a set of $N/2$ top and $N/2$ QCD jets, where N is a large number. For the i-th jet, we assign the "target output" variable: $y_i = 1$ if it is a top jet, and $y_i = 0$ if it is a QCD jet. Training consists of adjusting the weights so that the actual outputs of the ANN Y_i correspond as close as possible to the target outputs y_i, across the training set. To quantify the error, we use the logarithmic loss variable

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(Y_i) + (1-y_i)\log(1-Y_i)\right]. \tag{3.3}$$

The goal of training is to choose weights that minimize this function. We use the back-propagation algorithm [43], combined with gradient-descent minimization. In its simplest version, the algorithm can be summarized as follows [44]:

1. Initialize the weights of each link to small random values.

[3] In Eq. (3.1) and below, repeated indices are always summed over.

2. Repeat until convergence of log-loss, for each input vector ϵ_i:

   - Forward: Compute the output of each neuron until the output layer is reached, that is

$$\epsilon_i \rightarrow h_i^{(1)} = f(W_{ij}^{(1)}\epsilon_j + b^{(1)}) \rightarrow h_i^{(2)} = f(W_{ij}^{(2)}h_j^{(1)} + b^{(2)}) \rightarrow Y = f(W_j^{(O)}h_j^{(2)}) \quad (3.4)$$

   - Backward: Adjust the weights of each neuron by propagating backward the error at the output using

$$\begin{aligned} \delta^{(O)} &= (y - Y)Y(1 - Y) \text{ and } \delta_k^{(l)} = h_k^{(l)}(1 - h_k^{(l)}) \sum_j W_{kj}^{(l-1)}\delta_j^{(l-1)} && (3.5) \\ W_k^{(0)} &\rightarrow W_k^{(0)} + \eta\delta^{(O)}h_k^{(2)} \\ W_{jk}^{(2)} &\rightarrow W_{jk}^{(2)} + \eta\delta_j^{(2)}h_k^{(1)} \\ W_{jk}^{(1)} &\rightarrow W_{jk}^{(1)} + \eta\delta_j^{(1)}\epsilon_k && (3.6) \end{aligned}$$

   where η is a small parameter called the learning rate.

We used several well-known tricks to make this algorithm more efficient. First, instead of updating the weights after each jet ϵ_i, we used what is known as batch gradient descent so that the update on the weights is only done after all the jets in the training set have been processed. In that scenario, the updates on the weights are an average of the individual updates caused by each jet. Moreover, to reduce the odds of getting stuck at local minima we add what is known as a "momentum" to the updates. This means that the weights at iteration t, W_{ij}^t, are still being pushed by the update from the previous iteration ΔW_{ij}^{t-1}, for example

$$W_{ij}^t \rightarrow W_{ij}^t + \eta\delta_i^{(l)}h_j^{(l-1)} + \alpha\Delta W_{ij}^{t-1} \quad (3.7)$$

where $\alpha \in (0, 1)$ is a fixed parameter.

A major concern in using ANN classifiers is over-fitting the network to the training data. Over-fitting is a common problem in machine learning, in which the training procedure produces a classifier that emphasizes random fluctuations in the training data set, as opposed to the underlying trend. An over-fitted classifier would achieve excellent performance on the training set, but this will not generalize well to data sets which were not part of the training set, rendering it useless. Many techniques for avoiding over-fitting have been proposed in the literature. However, over several experiments we found that it was easier to avoid over-fitting simply by using more training data and ensembling several neural networks together. To determine the size of the training set $N_{\rm tr}$ needed to saturate the learning of our neural network, we studied the performance of the trained network on a cross-validation set of 50000 top and QCD jets, as a function of $N_{\rm tr}$. For this analysis, the performance is characterized by the ROC AUC (area under the receiver operating characteristic curve) performance metric, which assigns a value of 0.5 to a random classifier and a value of 1.0 to a perfect classifier. As can be seen on Fig. 2, performance steadily

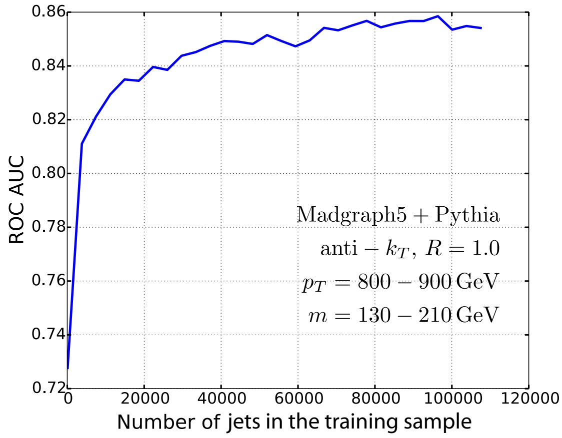

Figure 2. ROC AUC on a cross-validation set of 50 000 jets, vs. number of jets in the training set.

improves with the training set size until $N_{\mathrm{tr}} \approx 40000$ (*i.e.* 20000 top images and 20000 dijet images), after which convergence is achieved. This indicates minimal over-fitting beyond that point.

To further improve the performance of our tagger, we ensembled multiple neural networks together. The idea is to train B neural networks together, with the output given by the average of their outputs,

$$O = \frac{1}{B}\sum_{i=1}^{B} Y_i. \tag{3.8}$$

In our application, $B = 10$. All networks are trained using the same training set, but the jets are weighted. For the first network, all weights are set to one. Jets which are heavily misclassified by the first network are then assigned a larger weight, while jets which are correctly classified are assigned a smaller weight. This re-weighted training set is then used to train the second network, and so on. This procedure allows the training algorithm to focus on specific events that are particularly arduous to classify, improving overall performance. For some parameter choices, this method can be mapped to boosted methods such as ADAboost [45], where the weak classifiers are feed-forward ANNs.

4 Results

The ensemble of ANNs described above has been trained on sets of about 50,000 top and QCD jets each, in three p_T bins, 500 – 600 GeV, 800 – 900 GeV, and 1100 – 1200 GeV. It has then been applied to test sets consisting of about 15,000 top and QCD jets each, in the same p_T bins. The distribution of the neural network output O on the test sets is shown in Fig. 3. The classification power of this observable is clear from the figure: top jets are predominantly assigned $O \approx 1.0$, while QCD jets are predominantly assigned $O \approx 0.0$. To use the ANN ensemble as a top-tagger, we simply choose a threshold value O_{th}, and assign the “top tag” to any jet with $O \geq O_{\mathrm{th}}$ and the “QCD tag” to any jet with $O < O_{\mathrm{th}}$.

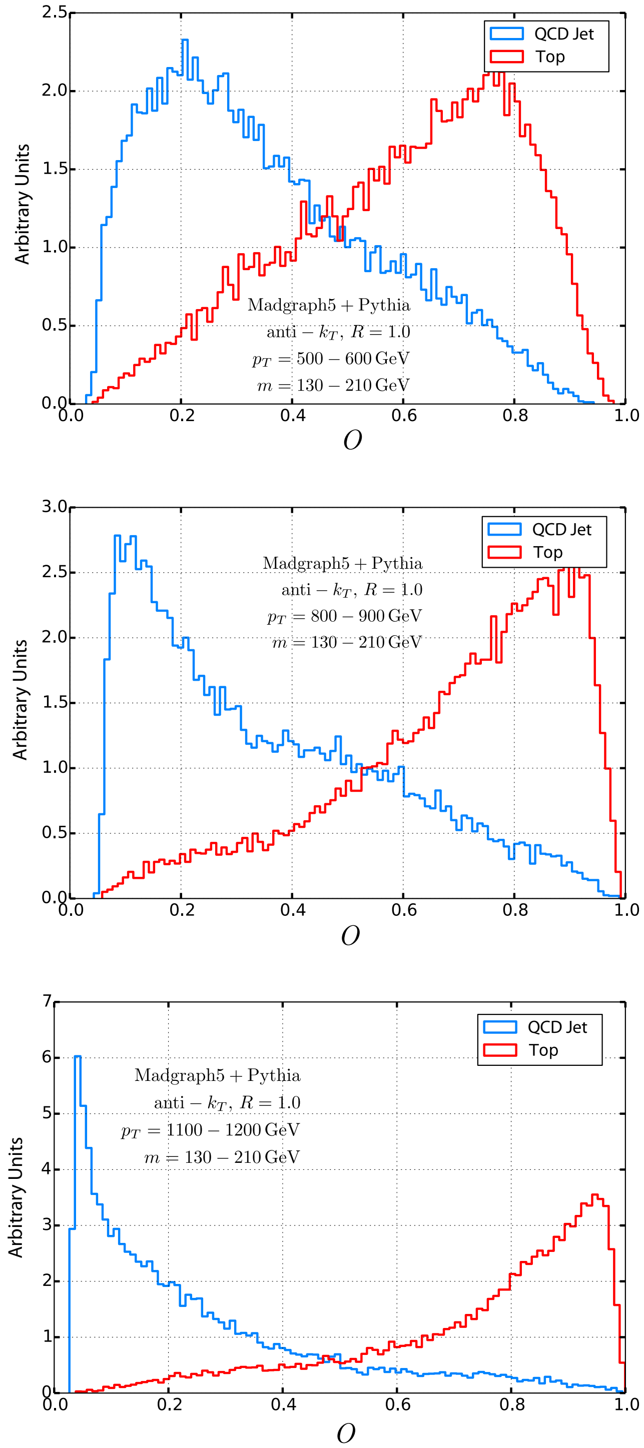

Figure 3. Distributions of the ANN output O on top (red) and QCD (blue) jet samples in three representative p_T ranges. All distributions are normalized to unit area.

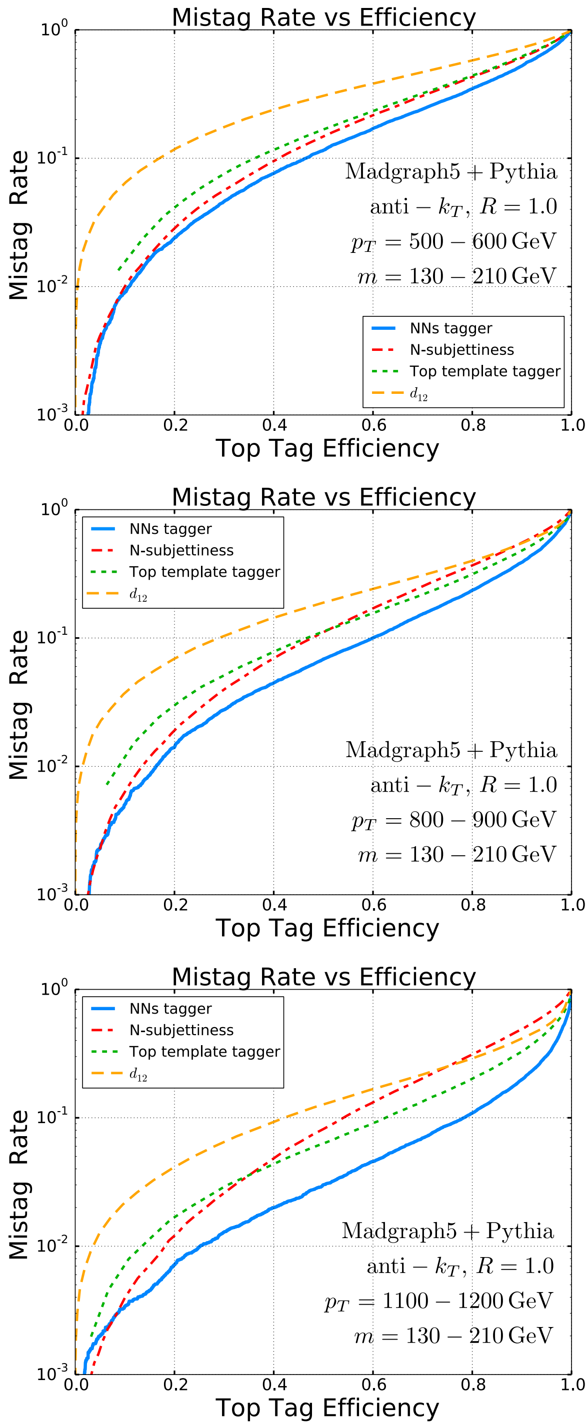

Figure 4. Efficiency vs. Mis-tag rate curves for the ANN tagger (blue/solid lines), for jets in three representative p_T ranges. For comparison, corresponding curves for three existing top taggers are also shown: d_{12} tagger (yellow/dashed), top template tagger (green/dotted), and N-subjettiness (red/dash-dotted).

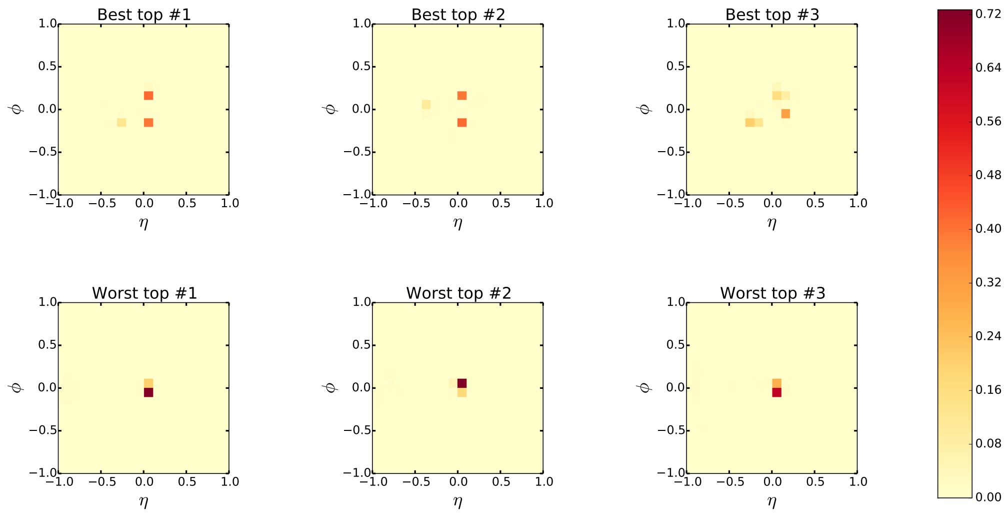

Figure 5. Energy deposit patterns for three jets with the highest (top row) and lowest (bottom row) ANN scores in the top sample with $p_T \in [800, 900]$ GeV.

To discuss the performance of the ANN tagger, it is convenient to define efficiency and mis-tag rates as follows:

$$\text{Eff} = \frac{N_{\text{top}}^{\text{top}}}{\mathcal{N}_{\text{top}}}, \quad \text{Mistag} = \frac{N_{\text{QCD}}^{\text{top}}}{\mathcal{N}_{\text{QCD}}}, \tag{4.1}$$

where $\mathcal{N}_{\text{top}}$ and $\mathcal{N}_{\text{QCD}}$ are the total number of jets in the top and QCD jet samples, respectively, and N_a^b is the number of jets in sample a tagged as jets of type b (a, b =top, QCD). Efficiency and mis-tag rates can be varied by varying the threshold $\mathcal{O}_{\text{th}}$. The performance of the ANN tagger is shown in Fig. 4, where for comparison we also show the performance of three representative existing taggers, described in the Appendix. In all cases, the ANN tagger outperforms the existing taggers, achieving lower mis-tag rates for the same tagging efficiency. The improvement is especially dramatic for high jet p_T: for example, for jets with $p_T \in [1.1, 1.2]$ TeV range, the ANN tagger achieves 60% tagging efficiency with about 4% mis-tag rate, about a factor of 2 lower than the best of the existing taggers in our comparison pool. This clearly demonstrates the promise of the ANN-based approach.

What physical features of the jet are identified by the ANN as the primary characteristics of a top jet? Some insight is provided by the energy deposit patterns of the highest-scoring and lowest-scoring jets, according to the ANN output $\mathcal{O}$, in the top sample. These are shown in Fig. 5. It is clear that the jets receiving high scores are characterized by well-defined three-prong structure, with each of the three quarks from top decay forming a well-defined, relatively isolated subjet. The lowest-scoring jets are those where either the quarks are nearly collinear, or one of them is much softer than the other two (in

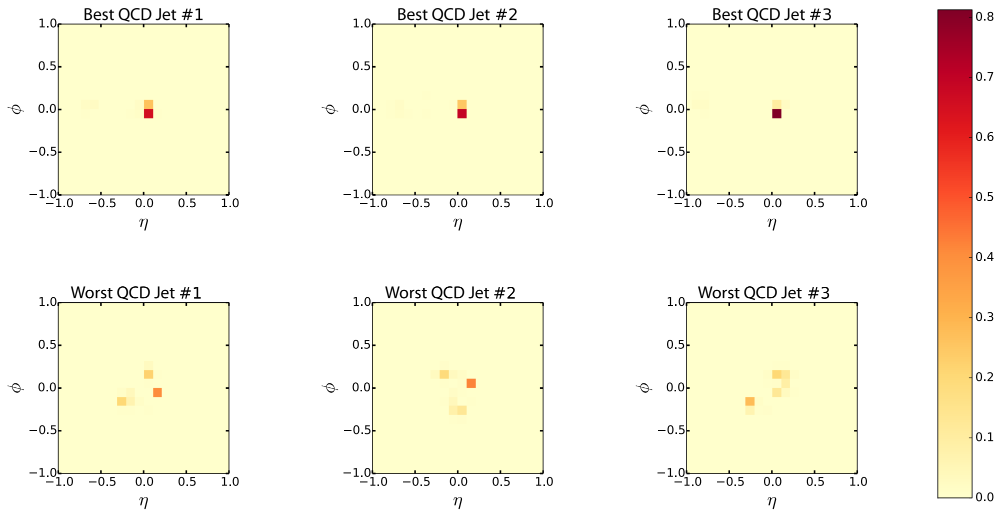

Figure 6. Energy deposit patterns for three jets with the lowest (top row) and highest (bottom row) ANN scores in the QCD jet sample with $p_T \in [800, 900]$ GeV.

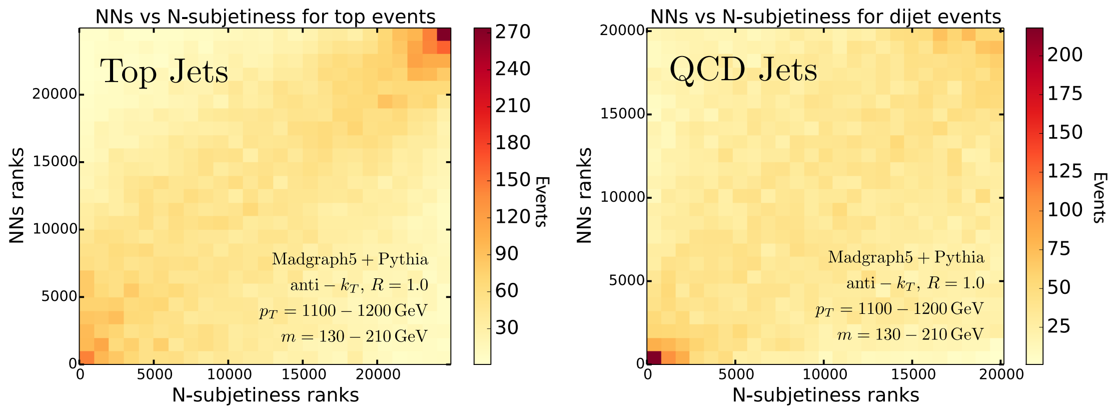

Figure 7. Correlation between the rankings of jets according to N-subjettiness (horizontal axis) and ANN score (vertical axis). Left: top sample, $p_T \in [1100, 1200]$ GeV. Right: QCD jet sample, same p_T range. Jets are ranked in order of increasing "topness" for both samples.

the detector frame). Likewise, the QCD jets receiving the highest scores, and thus most likely to be mis-identified as tops, have well-defined, isolated subjets, while the QCD jets correctly tagged as such do not: see Fig. 6.

To gain further insight, we studied correlations of the ANN scores with other observables used to tag tops. Table 1 contains the correlation coefficients between the ANN score and the output of the other taggers in our comparison pool, on a variety of samples used in our analysis. (The correlation coefficients are normalized so that 1.0 indicates perfect correlation and -1.0 perfect anti-correlation, while 0 indicates absence of correlation.) In

Tagger	Top		Dijet	
	$p_T \in [500, 600]$	$p_T \in [1100, 1200]$	$p_T \in [500, 600]$	$p_T \in [1100, 1200]$
TOM	0.50	0.52	0.52	0.65
N-sub.	0.59	0.52	0.48	0.31
ATLAS	0.33	0.44	0.42	0.72

Table 1. Correlation coefficients between the ANN score and the output of alternative taggers, in a variety of samples.

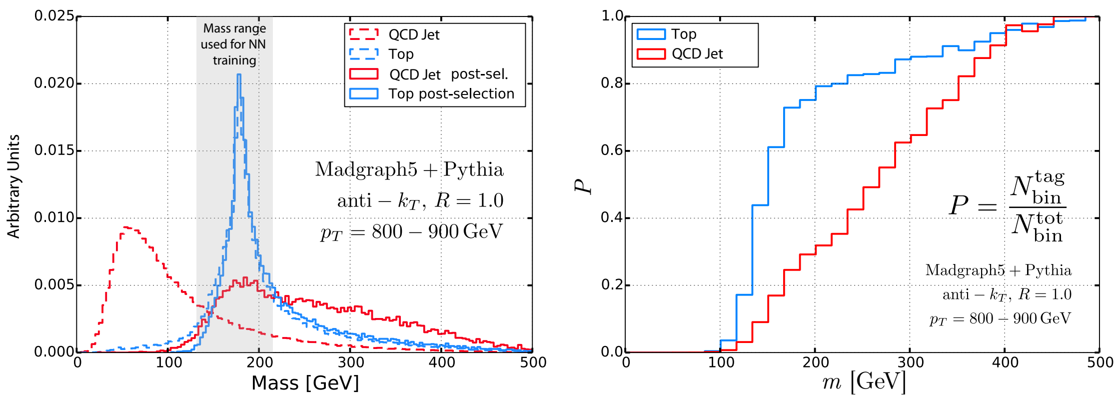

Figure 8. Left: Jet mass distributions for top (blue) and dijet (red) samples with $p_T \in [800, 900]$ GeV window, and no mass cut. Dashed lines: all jets; solid lines: jets tagged as tops by the ANN tagger. All distributions are normalized to unit total area. Right: probabilities for a jet in the top (blue) and dijet (red) samples to be tagged as a top jet by the ANN tagger.

all cases, we observe significant, though far from perfect, positive correlations, with coefficients ranging from about 0.3 to 0.7. A visual illustration is provided by Fig. 7, which shows that the ranking of jets according to the ANN score and the N-subjettiness are indeed correlated, in both top and light-jet samples; correlation plots for all other taggers and p_T ranges look very similar. This should not be surprising since all top taggers to some extent exploit the same physical characteristics of the boosted top jets. Nevertheless, as noted above, ANN systematically outperforms the other taggers in terms of tagging efficiency vs. mistag rates, indicating that the complicated non-linear observable created by the ANN learning process captures the information present in the jet substructure in a more optimal way. In other words, it seems that all taggers find roughly the same subset of jets to be "easily classifiable", and all have a very good success rate on this subset. However, the ANN tagger seems to be able to correctly classify a higher fraction of the jets outside of this subset, leading to higher overall success rate.

Another interesting question is how the ANN performance varies with the jet mass. The training samples and test samples in all plots shown so far only contain jets in a 130 . . . 210 GeV mass window, where most top jets are expected to lie. We also applied

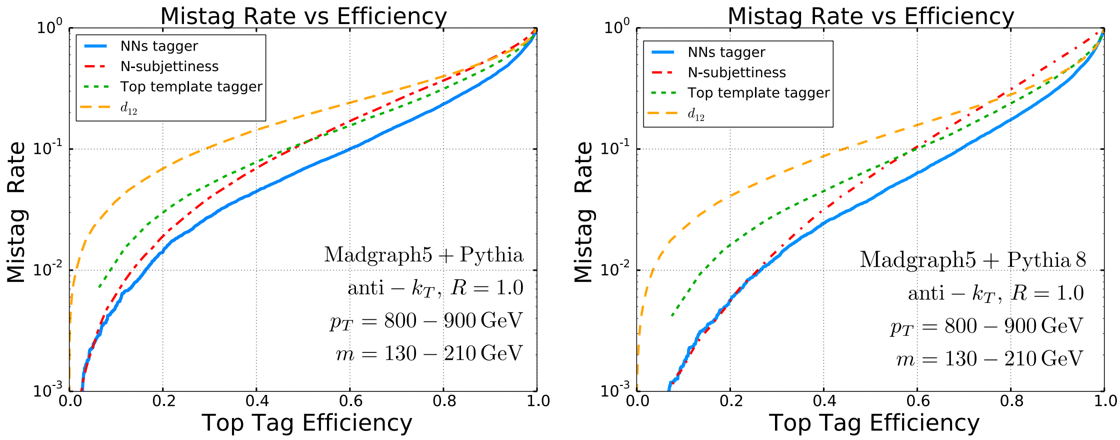

Figure 9. Efficiency vs. Mis-tag rate curves for the ANN tagger (blue/solid lines), on jet samples generated with `Pythia 6` (left) and `Pythia 8` (right). For comparison, corresponding curves for three existing top taggers are also shown: d_{12} tagger (yellow/dashed), top template tagger (green/dotted), and N-subjettiness (red/dash-dotted).

the ANN tagger to the full sample of jets in the [800, 900] GeV p_T range, without the mass cut. The jet mass distributions in this sample, before and after the ANN tagger is applied, as well as the tagging probability as a function of the jet mass, are shown in Fig. 8. (The cut on the ANN output used in the figure corresponds to the overall tag efficiency in the 130 . . . 210 GeV mass window of 70%.) For jet mass below 130 GeV, the probability of a positive top tag drops rapidly, for both top and QCD jets. This is presumably due to the fact that jets with a clear three-prong structure are unlikely to have a low mass. On the other hand, for jet mass above 210 GeV, the probability of a positive top tag is roughly independent of the jet mass. It should also be noted that the tag probability is smooth on the boundaries of the mass window selected for training, indicating that there is no strong dependence on the choice of the training sample. The ability of the ANN tagger to reject jets with low invariant mass may be useful in reducing effects of the pile-up.

The final issue we address is the IR-safety of the ANN output. As any observable in jet physics, the ANN score must be IR-safe (or at least Sudakov-safe [46]) to be useful. Canonically, IR-safety simply requires that the observable be unchanged by exactly collinear $1 \to 2$ patron splitting, or an emission of an infinitely soft gluon. Since neither process affects the energy deposits in calorimeter cells ε_{ab}, and since those energy deposits are the only information used by the ANN, its output is manifestly IR-safe by this definition. As a practical matter, however, one might still worry about the sensitivity of the output to non-perturbative physics involved in splittings at small, but finite, angles, and emission of gluons with small, but finite, energy. The modeling of this physics in MC generators such as `Pythia` involves approximations with poorly understood systematic errors, and if the ANN output were determined predominantly by features that depend strongly on the showering model, MC studies would clearly be of very limited utility in assessing the ANN performance on real data. To address this concern, we applied the

ANN tagger, trained as described above on jet samples showered with `Pythia 6`, to alternative jet samples generated with the same physics inputs but showered with `Pythia 8`. Showering algorithm of `Pythia 8` differs significantly from `Pythia 6`, in that it incorporates a p_T ordered showering as well as an increased number of underlying event modes and the capability to consider two hard processes. We also applied the three taggers in our comparison pool to the same sample. The result is shown in Fig. 9. The ANN tagger continues to perform well on test samples generated with a showering model different from the one used in the training set. This indicates that the features ANN uses to classify jets are physical, rather than artifacts of a particular showering model. Moreover, while there is a non-trivial dependence of the efficiency/mis-tag rate curves on the generator, the effect is of the same size for all taggers considered here. In other words, ANN does not appear to be unusually sensitive in this regard.

5 Discussion

In this paper, we proposed and explored a new approach to the analysis of jet substructure, specifically top-jet tagging, based on Artificial Neural Network (ANN). The main result of the analysis is captured in Fig. 4: the ANN tagger significantly outperforms traditional taggers on the MC "datasets" used in our study. In a sense, this should not come as a surprise: while the ANN uses the same input information as any other tagger, the training procedure constructs a non-linear function of these inputs which is specifically chosen to maximize its power to classify jets. This maximization takes place on a restricted but extremely broad set of functions, encoded in Fig. 1 or Eq. (3.1), and the resulting observable is probably not far away from the theoretical upper limit on classification performance. If this is indeed the case, the ANN can be useful in theoretical studies, serving as a benchmark for other observables used for boosted top tagging.

Being the first study of this novel approach to top tagging, the analysis presented here does not yet fully capture the complexity of the problem in a realistic experimental environment. The very promising results of this analysis strongly motivate further explorations. Some of the important outstanding issues include:

- The jets were extracted from event samples including only leading-order SM processes, $t\bar{t}$ and dijet. Subleading processes need to be included. In spite of their smaller rate, they may have outsize effect on the tagger performance: for example, pure QCD processes with high multiplicity of partons in the final state can create "accidental substructure" [47, 48], and the ANN would need to learn to distinguish it from real top jets.

- Pile-up has not been included in our simulations. While many methods to reduce the effects of pile-up have been suggested [19, 20], their interaction with the ANN tagger needs to be explored.

- Before the method can be applied to real data, concerns about possible MC biases in training the ANN need to be addressed. A preliminary study of this issue suggests

that the features that determine the ANN output are not strongly MC-dependent, see Fig. 9. However, a more extensive study of this issue is needed, ideally using control/validation samples from real LHC data. In principle, it may even be possible to train the ANN directly on real data, assuming that sufficiently robust training samples can be extracted. This approach would entirely remove concerns about MC biases, and warrants further investigation.

We plan to address some of these issues in future work.

Another important direction is to further improve the tagger performance. A clear limitation of our tagger is that it only uses HCAL information. Other pieces of information are highly relevant for top tagging, the most obvious one being a sub-jet b-tag. This information can certainly be combined with the algorithm presented here to construct an even more powerful tagger. Also, the tagger presented here is based on a rather simple NN architecture and training procedure; more advanced techniques, such as using a convolutional neural network or pre-training the neural network with unsupervised techniques, may result in improved performance.

Finally, while in this paper we focused exclusively on tops, this approach can equally well be applied to other boosted-object jets, such as W and h. It would be interesting to see if performance improvements with respect to traditional taggers can also be achieved in those cases.

In summary, the novel approach to jet tagging based on pattern-recognition techniques, specifically Artificial Neural Networks, shows promise of significant improvements in tagger performance. While the analysis presented in this paper is only the first step, we hope that this approach will eventually become a useful tool in experimental searches for new physics.

Acknowledgements

The authors are grateful for conversations with Fabio Maltoni and Jesse Thaler. We acknowledge the support of the U.S. National Science Foundation through grants PHY-1316222 (MC and MP) and PHY-0844667 (MP). SL is supported in part by the National Research Foundation of Korea grant MEST No. 2012R1A2A2A01045722. MB is supported in part by the Belgian Federal Science Policy Office through the Interuniversity Attraction Pole P7/37. LGA's research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 604102 (HBP).

A A Brief Description of Existing Top Taggers

For the purpose of comparison of the ANN tagger to the existing algorithms, we have chosen three existing methods, each one exploiting a different approach to boosted top tagging. In the following list, we give a brief description of the algorithms and the parameters we use for the analysis, while we refer the reader to the references within for detailed discussions.

- **Template Overlap Method (TOM)**: TOM [23–26] is a jet substructure algorithm which aims to match the energy distribution of a fat jet to a partonic structure which models the decay of a heavy boosted particle. TOM algorithm proceeds by comparing libraries of kinematically allowed parton level decays of massive particles ("templates") to the energy distribution of a fat jet. The quality of a match is quantified by the overlap function Ov, which minimises the difference between the parton transverse momenta and the amount of p_T deposited in small angular regions around the template patrons ("template sub cones"). An $Ov \sim 1$ score signals a top like jet, while a $Ov \sim 0$ is characteristic of light QCD jets. Here we use the `TemplateTagger v.1.0` [49] implementation of the TOM algorithm.

  There are many ways generation of template libraries can be implemented. For simplicity and processing speed, here we consider templates at fixed total transverse momentum matched to the mid-point in each fat jet p_T bin of the event samples (*e.g.* 550 GeV for fat jet $p_T = 500 - 600$ GeV). We generate the template states using a sequential scan of 40 steps in η, ϕ over the angular region of $R = 1.0$ around the fat jet axis. We match the template libraries to the energy distribution of the fat jet using fixed template sub cones of size $r_3 = 0.1, 0.15, 0.2$ for template $p_T = 1150, 850, 550$ GeV respectively, while we allow for the template resolution parameter $\sigma_a = p_{T,a}/3$, where $p_{T,a}$ is the transverse momentum of an individual template parton.

- **N-subjettiness:** Perhaps the most notable example of a "prong" tagger is N-subjettiness [21, 22]. The algorithm is based on calculating moments τ_N, which serve as estimates of how well the jet energy distribution can be divided into N regions. The τ_N are calculated by minimizing the p_T weighted distances between calorimeter energy depositions and trial axes which divide the distribution into N regions, over the space of possible axis configurations. The N-subjettiness tagger used in our comparisons is the version publicly available on `HepForge`.[4]

  For the purpose of top tagging the most useful observable is typically the ratio τ_3/τ_2, where a high score means that a jet distribution is described better by a three prong configuration. Conversely, a low τ_3/τ_2 score is characteristic of two prong jets. Note that in the analysis of this paper we used the angular weight exponent $\beta = 1$ in calculations of τ_N moments, as suggested in Ref. [22].

- **ATLAS top tagger**: Jet clustering history can provide useful insight into jet substructure. A notable example is the ATLAS top tagger [50] which utilises the differences between the top and light jets in the last step of jet clustering. The observable ATLAS uses is d_{12}, the value of the k_T norm at the clustering step which goes from two subjects to one final jet. The d_{12} observable is sensitive to the dynamics of hard splittings within the fat jet. The highly asymmetric splittings of typical light jets tend to be characterised by low values of d_{12} with a distribution which falls off sharply with the increase in d_{12}, while we expect typical top jets to be characterised by $d_{12} \sim m_t^2/4$.

[4]See `http://fastjet.hepforge.org/contrib/contents/latest.html`

In addition to d_{12}, ATLAS also imposes a lower cut on the trimmed jet mass of $m_j > 130$ GeV. Unless otherwise noted, here we omit the lower mass cut as the data samples we use for comparison are already restricted to a jet mass window in Eq. (2.1).

References

[1] G. Perez, *Top quark theory and the new physics searches frontier*, *Phys.Scripta* **T158** (2013) 014008.

[2] K. Agashe, A. Belyaev, T. Krupovnickas, G. Perez, and J. Virzi, *LHC Signals from Warped Extra Dimensions*, *Phys.Rev.* **D77** (2008) 015003, [hep-ph/0612015].

[3] B. Lillie, L. Randall, and L.-T. Wang, *The Bulk RS KK-gluon at the LHC*, *JHEP* **0709** (2007) 074, [hep-ph/0701166].

[4] M. Perelstein and A. Spray, *Four boosted tops from a Regge gluon*, *JHEP* **1109** (2011) 008, [arXiv:1106.2171].

[5] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, *Stop Reconstruction with Tagged Tops*, *JHEP* **1010** (2010) 078, [arXiv:1006.2833].

[6] J. Berger, M. Perelstein, M. Saelim, and A. Spray, *Boosted Tops from Gluino Decays*, arXiv:1111.6594.

[7] A. Azatov, M. Salvarezza, M. Son, and M. Spannowsky, *Boosting Top Partner Searches in Composite Higgs Models*, *Phys.Rev.* **D89** (2014) 075001, [arXiv:1308.6601].

[8] T. Flacke, J. H. Kim, S. J. Lee, and S. H. Lim, *Constraints on composite quark partners from Higgs searches*, *JHEP* **1405** (2014) 123, [arXiv:1312.5316].

[9] M. Backović, G. Perez, T. Flacke, and S. J. Lee, *LHC Top Partner Searches Beyond the 2 TeV Mass Region*, arXiv:1409.0409.

[10] M. Backović, T. Flacke, J. H. Kim, and S. J. Lee, *Boosted Event Topologies from TeV Scale Light Quark Composite Partners*, arXiv:1410.8131.

[11] B. Gripaios, T. Mller, M. Parker, and D. Sutherland, *Search Strategies for Top Partners in Composite Higgs models*, *JHEP* **1408** (2014) 171, [arXiv:1406.5957].

[12] J. Reuter and M. Tonini, *Top Partner Discovery in the $T \to tZ$ channel at the LHC*, arXiv:1409.6962.

[13] A. Altheimer, S. Arora, L. Asquith, G. Brooijmans, J. Butterworth, *et. al.*, *Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks*, *J.Phys.* **G39** (2012) 063001, [arXiv:1201.0008].

[14] M. Jankowiak and A. J. Larkoski, *Jet Substructure Without Trees*, *JHEP* **1106** (2011) 057, [arXiv:1104.1646].

[15] J. Bjorken and S. J. Brodsky, *Statistical Model for electron-Positron Annihilation Into Hadrons*, *Phys.Rev.* **D1** (1970) 1416–1420.

[16] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, *JHEP* **0807** (2008) 092, [arXiv:0806.0023].

[17] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, *et. al.*, *Substructure of high-p_T Jets at the LHC*, *Phys.Rev.* **D79** (2009) 074017, [arXiv:0807.0234].

[18] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys.Rev.Lett.* **100** (2008) 242001, [arXiv:0802.2470].

[19] D. Krohn, J. Thaler, and L.-T. Wang, *Jet Trimming*, *JHEP* **1002** (2010) 084, [arXiv:0912.1342].

[20] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys.Rev.* **D81** (2010) 094023, [arXiv:0912.0033].

[21] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **1103** (2011) 015, [arXiv:1011.2268].

[22] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **1202** (2012) 093, [arXiv:1108.2701].

[23] L. G. Almeida, O. Erdogan, J. Juknevich, S. J. Lee, G. Perez, *et. al.*, *Three-particle templates for a boosted Higgs boson*, *Phys.Rev.* **D85** (2012) 114046, [arXiv:1112.1957].

[24] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman, and I. Sung, *Template Overlap Method for Massive Jets*, *Phys.Rev.* **D82** (2010) 054034, [arXiv:1006.2035].

[25] M. Backović, J. Juknevich, and G. Perez, *Boosting the Standard Model Higgs Signal with the Template Overlap Method*, *JHEP* **1307** (2013) 114, [arXiv:1212.2977].

[26] M. Backović, O. Gabizon, J. Juknevich, G. Perez, and Y. Soreq, *Measuring boosted tops in semi-leptonic $t\bar{t}$ events for the standard model and beyond*, *JHEP* **1404** (2014) 176, [arXiv:1311.2962].

[27] **D0 Collaboration** Collaboration, V. Abazov *et. al.*, *A precision measurement of the mass of the top quark*, *Nature* **429** (2004) 638–642, [hep-ex/0406031].

[28] P. Artoisenet, V. Lemaitre, F. Maltoni, and O. Mattelaer, *Automation of the matrix element reweighting method*, *JHEP* **1012** (2010) 068, [arXiv:1007.3300].

[29] D. E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, *Phys.Rev.* **D84** (2011) 074002, [arXiv:1102.3480].

[30] D. E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, arXiv:1211.3140.

[31] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop*, *JHEP* **1405** (2014) 146, [arXiv:1402.2657].

[32] **ATLAS Collaboration** Collaboration, G. Aad *et. al.*, *A search for $t\bar{t}$ resonances in lepton+jets events with highly boosted top quarks collected in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *JHEP* **1209** (2012) 041, [arXiv:1207.2409].

[33] **ATLAS Collaboration** Collaboration, G. Aad *et. al.*, *Search for resonances decaying into top-quark pairs using fully hadronic decays in pp collisions with ATLAS at sqrt(s) = 7 TeV*, arXiv:1211.2202.

[34] **CMS Collaboration** Collaboration, *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging, CMS-PAS-JME-09-001*, .

[35] **CMS Collaboration** Collaboration, *Jet Substructure Algorithms, CMS-PAS-JME-10-013*, .

[36] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, arXiv:1407.5675.

[37] F. Maltoni and T. Stelzer, *MadEvent: Automatic event generation with MadGraph*, *JHEP* **0302** (2003) 027, [hep-ph/0208156].

[38] T. Sjostrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **0605** (2006) 026, [hep-ph/0603175].

[39] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput.Phys.Commun.* **178** (2008) 852–867, [arXiv:0710.3820].

[40] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, *Eur.Phys.J.* **C72** (2012) 1896, [arXiv:1111.6097].

[41] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, *JHEP* **0804** (2008) 063, [arXiv:0802.1189].

[42] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1st ed., 1996.

[43] P. Werbos, *Beyond regression: new tools for prediction and analysis in the behavioral sciences.*, *Thesis (Ph. D.), Harvard University* (1975).

[44] S. Marsland, *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, 1st ed., 2009.

[45] Y. Freund and R. E. Schapire, *A short introduction to boosting*, in *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 1401–1406, Morgan Kaufmann, 1999.

[46] A. J. Larkoski and J. Thaler, *Unsafe but Calculable: Ratios of Angularities in Perturbative QCD*, *JHEP* **1309** (2013) 137, [arXiv:1307.1699].

[47] A. Hook, E. Izaguirre, M. Lisanti, and J. G. Wacker, *High Multiplicity Searches at the LHC Using Jet Masses*, *Phys.Rev.* **D85** (2012) 055029, [arXiv:1202.0558].

[48] T. Cohen, E. Izaguirre, M. Lisanti, and H. K. Lou, *Jet Substructure by Accident*, *JHEP* **1303** (2013) 161, [arXiv:1212.1456].

[49] M. Backović and J. Juknevich, *TemplateTagger v1.0.0: A Template Matching Tool for Jet Substructure*, arXiv:1212.2978.

[50] **ATLAS Collaboration** Collaboration, G. Aad *et. al.*, *A search for $t\bar{t}$ resonances in the lepton plus jets final state with ATLAS using 4.7 fb^{-1} of pp collisions at $\sqrt{s} = 7$ TeV*, *Phys.Rev.* **D88** (2013) 012004, [arXiv:1305.2756].